DOCUMENT RESUME

ED 161 425 IR 006 340

AUTHOR Vance, David
TITLE What Are Data? Museum Data Bank Research Report Number 1.
INSTITUTION Margaret Woodbury Strong Museum, Rochester, N.Y. Museum Data Bank Committee.
PUB DATE Nov 74
NOTE 13p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Automatic Indexing; *Classification; *Data Bases; Information Retrieval; *Information Storage; *Museums; *Organization

ABSTRACT

This paper describes the process of automatic extraction of implicit--global--data from explicit information by file inversion and threading. Each datum is the symbolic representation of a proposition, and as such has a number of movable parts corresponding to the ideal elements of the proposition represented; e.g., subject, predicate. A third element is expressed by word order--the relationship between subject and predicate. Appropriate relationships are: is included in; includes; intersects; and excludes. The predicate is the "classifier" of the subject. Inversion consists of turning data around, putting the subject in the place of the predicate and some permutation of the predicate in place of the subject, while changing the sense of the relation--normally from "is included in" to "includes." In this way the series of data with the same subject, instead of describing a single entity, becomes the inventory of a broad class, bringing together things with common attributes. Threading forms a data organization where the subject of each datum after the first is identical with the term in the predicate of at least one of the preceding data. This can become a powerful research tool. (Author/MBR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161425

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

MUSEUM DATA BANK RESEARCH REPORT

Number 1

What Are Data?

by

David Vance

November 1974

This report, or any portion thereof, may be reproduced for any non-commericial purpose provided appropriate credit is given to the author and the Museum Data Bank Committee as publisher.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. G. Chenhall

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Museum Data Bank Committee
Robert G. Chenhall, Chairman

Margaret Woodbury Strong Museum
700 Allen Creek Road
Rochester, N.Y.

IR000340

Forward

There are two kinds of information in a data base, information that is explicitly stated and information that is implicit or "global." In a museum inventory, the fact that a specimen was collected at certain map coordinates in the Caribbean Sea or that it was worn by General U. S. Grant or that it was painted by Rembrandt is normally explicit amd may be read directly. The facts that a certain number of other specimens share the same attributes and that these may be found in certain places are normally implicit and may be obtained by counting, sorting or otherwise manipulating discrete elements of the data base.

The best reason for recording data in machine-readable form is to automate these manipulations. Thus transcription, of specimen data only, gives the user potential access to implicit information as well.

The first step in the extraction of implicit information is indexing but there are two ways to index. The first is the familiar index to key expressions. Such an index has no information content except that certain records contain certain expressions and so refer <u>in some way</u> to the concepts those expressions represent. It does not advance the extraction of implicit information.

The second kind of index points to <u>propositions</u> and itself contains information. If complete, it may actually carry the entire information content of the indexed data base, with this difference: what was originally explicit is here implicit while much that was implicit is here explicit. This step is called "file inversion."

Other implicit information may be made explicit by statistical analysis of (explicit or implicit) numerical data. Yet other implicit information can be extracted through syllogistic manipulation of original or inverted data. And, finally, a process called "threading," discussed below, may reveal even more.

All these processes rely upon some variation of a data organization called "tagged descriptor" or "tag and value" format, without which it is impossible to index or even identify propositions among the data. The following essay presents a view of data that elucidates the automatic extraction of implicit from explicit information and demonstrates the necessity for a tagged descriptor format or its logical equivalent.

Data as Propositions

It is the thesis of this paper that a datum is the symbolic representation of a proposition, which is asserted, denied or assigned a numerical probability.

A proposition must be about something; and information about something can be conveyed only by relating it to something else, presumed to be known. By analogy to grammar, a proposition is about an unknown aspect of a "subject," which is related to a known "predicate." By analogy to theory of equations, a proposition is about an unknown quantity, which is related to an expression that can be assigned a value.

Either subject or predicate may be replaced by an equivalent expression, in which an unlimited number of other things may be referenced. Moreover, a proposition can be transformed in a manner analogous to the transformation of an algebraic equation, so that the name of any one thing may stand alone as the unknown subject, all other terms being relegated to the predicate.

Thus no one entity is inherently the subject of a propostition. The subject is whichever of two or more entities named is currently considered the unknown, i.e. the name for which the proposition is to be "solved." Solution, in this context, consists in restricting the reference of the subject term by assigning it to a class specified by the predicate.

Elements of a Datum

A datum, as the symbolic representation of a proposition, must have a number of movable parts corresponding to the ideal elements of the proposition represented, e.g.

⟨subject⟩ ⟨predicate⟩

A third element, expressed above by word order, is the relationship between subject and predicate. Since it is not a law of nature that subjects precede predicates, a better representation is the following:

$\langle \text{entity}_1 \rangle$ ⟨relationship⟩ $\langle \text{entity}_2 \rangle$

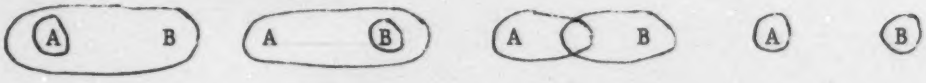
If we consider that entity_1 is the subject, then it must simply be named. The set of all data with the same name in the position of entity_1 constitutes a description of the named entity. In the context of museum data, entity_1 is often but not necessarily an accession number, a site number, a personal name, a place name or a taxon name. Any of these, alone or in combination with other names, may also figure in entity_2; but the same name should not appear in both entities for the same reason that the unknown whould not be found on both sides of an algebraic equation. Words

classified grammatically as "verb" or "adjective" are here considered names and no distinction is made between "proper" and other nouns.

Relationships, in this context, are not to be confused with the "tags" often used in museum data banking. If propositions are to serve as a vehicle for calculation, the middle term must have the character of a mathematical relation analogous to the "=," " > ", etc. of algebra or the " <=> " (if and only if), "&," etc. of logical notation. It is suggested that the appropriate set of relations includes the following:

1. is included in
2. includes
3. intersects
4. excludes

represented by the Venn diagrams:



1 2 3 4

and that the stored form of data in museum catalogs as we know them uses only the first of these. It is never expressed but invariably assumed.

The often complex ideas represented by our tags (or the positional equivalent) must be considered parts of the even more complex expression designating $entity_2$, the predicate. Thus in MCN* notation (Vance 1973):

oak 50==

is understood to refer to a subject named elsewhere within the same record and to mean:

subject / is included in / the set of objects made, at least in part, of oak (wood)

In this analysis, the subject and predicate are regarded as the designations of two sets, each with at least one member, such that all members (normally one) of the subject set are also members of the predicate set. If the predicate names a set of only one member, identity is asserted and the two terms are synonymous.

---

*Museum Computer Network, Inc., a non-profit corporation with museums and similar non-profit organizations as members, devoted to cooperative development of computer applications in museology.

The foregoing is trivial in the sense that it is logically equivalent to the traditional view of entity$_1$ as an "object" and entity$_2$ as its "attribute" (Lewis 1965), "modifier" or "descriptor." It is, likewise, equivalent to viewing the predicate as a "classifier" of the subject. In this last view the list of n propositions with the same subject defines that subject as the intersection of n classes.

The sole advantage sought here in emphasizing the propositional nature of data is to show the need for elements that are not always expressed. It is the author's position that if any one of these elements be truly lost - neither expressed nor implied - no information can be communicated and no data said to exist.

In practice, data are often represented by a succession of comparable numbers or other character strings without associated representation of subjects, relations or even tags. The values expressed are simply the variable elements of a series of data. It is understood that elements not expressed hold constant or change in a regular sequence. The practice gives rise to a common usage in which the word "data" (always in plural form but treated grammatically as a substance, like milk or honey) refers only to the mass of expressed variables. The author has no quarrel with this usage since it leads to no error. The fact remains that data, in this sense of the word, is (sic) meaningless in itself. In conjunction with the understood context, each element in the series asserts a meaningful proposition. It becomes a datum.

The list of elements is not yet complete for little has been said of entity$_2$, the set designated by the predicate. In all the tag-and-value notations used in museum data banking the predicate consists of a tag and descriptor conjoined. Consider a second example, again in MCN notation:

25 52==

Here, too, the subject has been named elsewhere in the record and the relation "is included in" is understood. An English paraphrase is:

> subject / is included in / the set of objects weighing 25 grams, to the nearest gram

The agreed meaning of "52==" is:

> the set of objects weighing _____ grams, to the nearest gram

where the blank requires a positive decimal integer as the variable part of the predicate. In the earlier example the meaning of "50==" is:

> the set of objects made at least in part of _____

where the blank requires the designation of a material. "Oak" is understood to mean oak wood. "Oaken" would be an adequate alternative though its adjectival form would call for revision of the English paraphrase.

These examples make it clear that the distinction between tag and "value" is not natural but arbitrary - a matter of convenience. The more information is packed into the agreed meanings of tags, the more specific the tags become and the more tags must be defined. As the number and specificity of tags in a system increases, the mean number of values appropriate for use with each tag (its "authority table") decreases.

A tag's information content is coded because its natural language paraphrase is relatively long, susceptible of expression in many, equally satisfactory ways, and utterly invariable in meaning. Indeed, it is an open question whether a perfectly complete and accurate natural language paraphrase can be found for any tag not expressing a tautological relationship such as synonymity. Certainly the author has yet to find one.

Values may, of course, be coded too; but in most present data banks they are not because values are relatively short, easily normalized and often required to sort lexically. At least one functioning museum data bank, at the Museo Nacional de Antropologia in Mexico City, employs a fixed code to represent the entire predicate without separately expressed tags and values. (Castillo Tejero et al. 1972)

Thus far, the following separately expressable elements of a datum have been distinguished:

⟨probability⟩ ⟨ [subject] [relation] [(tag) (value)] ⟩

Since this paper is not to reach book length, it will be assumed henceforth that the probability and relation remain constant. The probability will be considered equal to 1, categorical assertion; and the relation will be that of inclusion of the subject entity in the predicate entity.

Elements of a Predicate

Before examining the various manipulations of the remaining elements (subject, tag and value) it is necessary to consider the possibility of associating multiple values with a given tag as well as the related concept of "sub-descriptors."

Once again using MCN notation, since it is typical of data representation by tag and value, consider the following:

van Rijn, Rembrandt H. 70==

An English paraphrase is:

subject / is included in / the set
of works of art attributed to
Rembrandt H. van Rijn

but the same proposition can as well be expressed in a hypothetical system wherein a tag, say "p==," signifies only that the associated value is a

personal name. However, with less information supplied by the tag, more must be given by the value(s) to express the same proposition, as follows:

attribution: van Rijn, Rembrandt H. p==

The meaning is unchanged but the meaning of the predicate in the latter example is distributed over three terms instead of two.

This change, in form only, becomes significant when the two data are "inverted" for indexing, one of the manipulations to be discussed below. The MCN predicate:

van Rijn, Rembrandt H. 70==

as an index entry, can be sorted in only two significant ways, under "van Rijn" and under "70." The former brings together all data concerning the man and, under that, segregates data in which his name is given as the attribution of a work of art. The other significant order brings together all attribution data and segregates them by artist.

The hypothetical predicate:

attribution: van Rijn, Rembrandt H. p==

can be indexed in six potentially significant ways, corresponding to the six permutations of three elements.

The permutation:

van Rijn, Rembrandt H./attribution/ p==

brings together all references to van Rijn and, among these, segregates attribution data. The second permutation:

van Rijn, Rembrandt H./p==/attribution

does exactly the same thing. The source of this redundancy lies in the datum for "van Rijn, Rembrandt H." never refers to an entity that is not a person. The permutation:

attribution/vanRijn, Rembrandt H./p==

brings together all attribution data and, among these, segregates attributions to van Rijn. The permutation:

attribution/p==/van Rijn, Rembrandt H.

brings together all attribution data, among these segregates attributions to persons (as opposed to schools, workshops, tribes and cultures) and, among these, further segregates attributions to van Rijn. The permutation:

p==/van Rijn, Rembrandt H./attribution

brings together all data about persons, among these, segregates data

about van Rijn and, among these, further segregates attribution data. Finally, the permutation:

p==/attribution/van Rijn, Rembrandt H.

brings together all data about persons, among these, segregates attribution data and, among the latter, further segregates attributions to van Rijn.

The foregoing illustrates two rules bearing upon the use of multiple, separately significant terms within the predicate:

1. The number of potentially useful index references to a datum equals the maximim permutations of the number of separately significant terms in its predicate.

2. Redundancy among the separately significant terms of the predicate results in two or more permutations with the same meaning.

In both rules the words "separately significant" are of critical importance. They refer to terms that are permutable and equal in rank. No one term of the predicate illustrated is inherently the qualifier of one of the others. The terms of the predicate are, as it were, "multiplied" together. This is a fundamental difference between the above and the following example.

Here a predicate, superficially similar to the last, is shown to be quite different in structure, i.e. in the relationships among its component parts. This example is in MCN notation. The subject, specified as usual elsewhere in the record, is the catalog of a museum exhibition; and the predicate illustrated specifies a place where that exhibition was shown:

U.S.A.; NY; New York; Museum of Modern Art, The 540==

This predicate has five distinct parts, four values and a coded tag; but there are only two, not 120, potentially useful permutations. The reason is that each value after the first is the qualifier of its predecessor. The meaning of the series of values depends upon their order and would be destroyed by permutation. In other words, the four apparent values are actually one, the full specification of one of the places where museum exhibitions are held. Predicate terms related in this way are, as it were, "added" together. The number of separately significant terms in this predicate is two and so is the maximum number of permutations.

Despite the shortage of useful permutations a predicate structured in this manner does convey more than the full specification of a place: it locates an institution (at least its exhibition facilities) with respect to a rigid hierarchy of place names. This structure, within the predicate of one datum, actually implies several other propositions, e.g.:

a place named "U.S.A."/is included in/the set of places that have a subdivision named "NY"

This datum and others in the same form are incidental, adding nothing to the description of the museum exhibition. Another series of implied data is important and accounts for the choice of predicate format. This series, to cite only one example, includes:

> subject / is included in / the set of
> catalogs of exhibitions held in New York,
> NY, U.S.A.

Functionally, the predicate last illustrated is our first example of one including or composed of sub-descriptors, here indicated by position rather than by sub-tags. The following illustration, with sub-tagged descriptors, is hypothetical since neither MCN notation nor any of the closely related systems supports this data format. It has been used elsewhere, however. (Gundlach 1970) The made up notation used below is derived from the MCN system and assumes that the subject, named elsewhere, is a work of art:

> van Rijn, Rembrandt H. (Dutch 208==) (+1606 212==)
> (+1669 214==) 70==

In this example "70==" is the familiar attribution tag and the sub-tags "208==," "212==" and "214==" label nationality, year of birth and year of death, respectively. The plus sign marks the dates as A.D.

Another plausible representation might dispense with a main entry (the artist's name) and incorporate sub-tags such as "N1==," "N2==" and "N3==" for first given name, second given name and family name, respectively:

> (Rembrandt N1==) (H. N2==) (van Rijn N3==)
> (Dutch 208==) (+1606 212==) (+1669 214==) 70==

In such a notation the order in which sub-descriptors are listed loses all importance. Alternatively, a rigidly specified sequence of sub-descriptors may take the place of sub-tags (Curran and Avran 1967). By the same logic sub-sub-descriptors can be introduced but an illustration won't be attempted.

It is possible but not rewarding to regard the last two examples above as data, two possible representations of a proposition to the effect that some work of art is a member of the set of works attributed to one of the artists named Rembrandt H. van Rijn, among these artists, to one of the Dutch ones, among these Dutch Rembrandts, one of those born in A.D. 1606, and among these, one of those who died in A.D. 1669!

This would obviously not reflect the intent of anyone recording data in the way illustrated. He assumes, without saying so, that there is only one Rembrandt H. van Rijn and means to express four separate propositions:

subject / is included in / the set of works of art attributed to Rembrandt H. van Rijn

Rembrandt H. van Rijn / is included in / the set of persons of Dutch nationality

Rembrandt H. van Rijn / is included in / the set of persons born in A.D. 1606

Rembrandt H. van Rijn / is included in / the set of persons deceased in A.D. 1669

and, in fact, these propositions have been expressed, though not in a form that facilitates the data manipulations described below, inversion and threading. Conversion of the given representation to four separate data (one about a work of art and three about a person) opens the way to these steps. It also procures a practical benefit since, whatever the number of Rembrandt's cataloged works, the three data about the artist need be banked only once - not once for each work.

Manipulations of Data

The preceding analysis has been aimed at data manipulation, systematic transformation intended to extract and represent propositions that are implicit in raw observations without being explicit or obvious. This process goes to the heart of both data banking and museology - insofar as museums are more than places of amusement.

The two fundamental organizations of data obtained through manipulation are inversion and threading. In form, both steps are as simple as addition. In practice both are as vulnerable to inaccurate givens as multiplication. Both operate upon the basic propositional form:

⟨subject⟩ ⟨relation⟩ ⟨predicate⟩

Inversion, in conjunction with sorting and statistical analysis, proceeds from the mass of direct observations, often of no interest in themselves, to general truths and, from these, sometimes, to insight into the past, into tendencies of change and into future probabilities. Such are the goals that justify collecting things and facts.

The mechanism of inversion consists in turning data around, putting the subject in place of the predicate and some permutation of the predicate in place of the subject, while changing the sense of the relation (normally from "is included in" to "includes.") Now the series of data with the same subject (assembled through the sorting process), instead of describing a single entity (e.g. a preserved specimen), becomes the inventory of a broad class, a class that is significant insofar as sig-

nificant data were collected in the first place. Such an inventory, or a shorter one resulting from the intersection of classes, brings together things with common attributes. Unexpected correlations among the contents of classes throw light upon the natural and social forces that left things as we find them. Often the factors uniting a class of specimens are time and place of origin. The whole set of available things that took their present forms under the same conditions tends to reveal what those conditions were. We have no other way of knowing.

The threading process forms another kind of data organization, a series of data, in their original form, inverted or of both kinds. The defining characteristic of this series is that the subject of each datum after the first is identical with a term in the predicate of at least one of the preceding data. The threading process may be carried to any length, subject to the rule that no one datum appear twice in the same thread, a condition that closes a loop of endless repetition. A short thread is illustrated in the last examples above, where a datum about a work of art includes in its predicate the name of an artist. This is followed by three data with that name as their common subject. One of these contains in its predicate the word "Dutch", the adjectival designation of The Netherlands. The thread might continue with a series of data about that country.

Techniques for control of threading are only beginning to be understood and data bases capable of supporting it have only begun to appear. The characteristic of such data bases is that identical, normalized terms are systematically used in the subjects and predicates of different data.

Clearly, threading cannot produce the sweeping generalizations that are the objective of inversion. A thread is coherent but rambling and, pursued at random, will produce correct but useless "chains" in the style of Tristram Shandy. It seems certain, however, that under suitable program control threading can become a powerful research tool, for the method is the one we use in the library, following our sources from subject card to author card, thence to citation, and so to another author.

Possibly the algorithm for retrieving a significant thread will rely upon a predetermined cycle of tags to be sought in one or more specified sequences.

One elementary application of threading will be in the use of machine-readable thesauri. Threading is necessary, too, for the use of structured and joined records. More sophisticated applications, employing "citation search" methods, may enter the realm of basic research. One may speculate that controlled threading within a very complex data base may discover unsuspected networks of relationships, revealed by the more than random frequency of connections among a limited set of names. Such loci of related ideas may manifest themselves as local "entanglements," where a thread is interrupted repeatedly upon returning to a datum it has already

encountered. The significance of such a thicket of relationships may not be obvious. Sometimes it will raise a new problem to be investigated.

Afterword

This paper must end with a warning. Mathematical statements can be manipulated with confidence that true premises and legal transformations will yield true conclusions; but this is so only if the meanings of all terms are exactly known. Every element must be univocal. The same is true of data manipulation but the necessary univocal terms are not easily refined from the low grade ore of natural language.

There is therefore, an overriding need for definition and rigorous vocabulary control. Often the words one least suspects conceal traps. Does "France" include Martinique? Tahiti? Did it formerly include Algeria? How does the sense of this word change in a medieval context? Does it always include Burgundy - retroactively? What will be the consequences of calling Picasso Spanish but including him in the School of Paris?

REFERENCES CITED

Castillo Tejero, Noemí; Silvia Garza de G. y Román Piña Chán

1972 Diccionario de terminos basicos para describir y catalogar las colecciones arquelogicas del Museo Nacional de Antropología. Antropología Matematica 21.

Curran, Ann T. and Henriette D. Avram

1967 The Identification of Data Elements in Bibliographic Records. Final report of the Special Project on Data Elements for the Subcommittee on Machine Input Records (SC-2) of the Sectional Committee on Library Work and Documentation (Z-39) of the United States of America Standards Institute.

Gundlach, Rolf

1970 Grunlagen u. Strukturen der Konponentendeskription in der historischen Dokumentation. Nachrichten fur Dokumentation, Beiheft Nr. 20.

Lewis, Geoffrey D.

1965 Obtaining information from museum collections and thoughts on a national index. Museums Journal 65, no. 1.

Vance, David

1973 Computers in the Museum. IBM application manual (GE20-0406-0).